Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W633–W638***
doi:10.1093/nar/gkm350

INDELSCAN: a web server for comparative identification of species-specific and non-species-specific insertion/deletion events

Feng-Chi Chen[1], Chueng-Jong Chen[2] and Trees-Juen Chuang[2,*]

[1]Division of Biostatistics and Bioinformatics, National Health Research Institute, Miaoli County 350, Taiwan and [2]Genomics Research Center, Academia Sinica, Taipei 11529, Taiwan

Received January 30, 2007; Revised April 12, 2007; Accepted April 23, 2007

ABSTRACT

Insertion and deletion (indel) events usually have dramatic effects on genome structure and gene function. Species-specific indels have been demonstrated to be associated with species-unique traits. Currently, indel identifications mainly rely on pair-wise sequence alignments (the 'pair-wise indels'), which suffer lack of discrimination of species specificity and insertion versus deletion. Also, there is no freely accessible web server for genome-wide identification of indels. Therefore, we develop a web server—INDELSCAN— to identify four types of indels using multiple sequence alignments that include sequences from one target, one subject and $\geq$1 out-group species. The four types of indels identified encompass target species-specific, subject species-specific, non-species-specific and target-subject pair-wise indels. Insertions and deletions are discriminated with reference to out-group sequences. The genomic locations (5′UTR, intron, CDS, 3′UTR and intergenic region) of these indels are also provided for functional analysis. INDELSCAN provides genomic sequences and gene annotations from a wide spectrum of taxa for users to select from, including nine target species (human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*), opossum (*Monodelphis domestica*), chicken (*Gallus gallus*), zebrafish (*Danio rerio*), fly (*Drosophila melanogaster*) and yeast (*Saccharomyces cerevisiae*) and >35 subject/out-group species, ranging from yeasts to mammals. The server also provides analytic figures and supports indel identification from user-uploaded alignments/annotations. INDELSCAN is freely accessible at http://indelscan.genomics.sinica.edu.tw/IndelScan/.

INTRODUCTION

Insertions and deletions (indels) represent a major force of genome evolution. It has been revealed that indels occur at a surprisingly high frequency and contribute to sequence divergence more than nucleotide substitutions do (1–4) between very closely related species, such as human and chimpanzee (3,5–7). Indels that occur after speciation (i.e. species-specific indels) can lead to significant changes in phenotype (8–10) in light of their dramatic effects on genome structure and gene function (6). Therefore, genome-wide analysis of species-specific indels and non-species-specific indels ('NSS' indels, i.e. indels of which species specificity is not observed) may shed some light on the mechanisms of genome evolution and functional divergence. However, currently no freely accessible web-based server is available for genome-wide identification of such indels. Hence, it is essential to develop a computational platform for this purpose.

Here we develop a web sever ('INDELSCAN') to infer species-specific and NSS indels using multiple sequence alignments from at least three species. The compared species should include one target species, one subject species and at least one out-group species. Note that the selection of out-group species is important for the resolution of INDELSCAN to infer insertions and deletions. The differentiation between insertions and deletions is evolutionarily important, and is usually impossible in pair-wise genome comparisons. In general, the out-group species should be more distantly related to both the target and subject species than they are to each other. Yet, an out-group very distant from both compared species may yield minimal resolution in the comparison. The identified target species-specific indels ('TSS' indels, i.e. indels that are specific to the target species), which are supported by out-group sequences, should have considerably higher accuracy than indels inferred from target-subject pair-wise comparisons if the subject genome remains a draft. Moreover, by incorporating annotations of the target genome, the

*To whom correspondence should be addressed. Tel: +886 2 27898730 348; Fax: +886 2 27898757; Email: trees@gate.sinica.edu.tw

Table 1. Currently available target, subject and out-group species at INDELSCAN (downloaded from the UCSC genome browser at http://hgdownload.cse.ucsc.edu/downloads.html.)

Target	Subject/out-group
Human (*Homo sapiens*)	Chimpanzee (*Pan troglodytes*), macaque (*Rhesus macaque*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), rabbit (*Oryctolagus cunicuhis*), dog (*Canis familiaris*), cow (*Bos Taurus*), armadillo (*Dasypus novemcinctus*), elephant (*Loxodonta Africana*), tenrec (*Echinops telfairi*), opossum (*Monodelphis domestica*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), tetraodon (*Tetraodon nigroviridis*) and Fugu (*Fugu rubripes*)
Mouse	Human, chimpanzee, macaque, rat, rabbit, dog, cow, armadillo, elephant, tenrec, opossum, chicken, frog, zebrafish, tetraodon and fugu
Rat	Human, mouse, dog, cow, opossum, chicken, frog and zebrafish
Dog	Human and mouse
Opossum	Human, mouse, rat, chicken, frog and zebrafish
Chicken	Human, mouse, rat, opossum, frog and zebrafish
Zebrafish	Human, mouse, opossum, tetraodon, fugu, and frog
Fly (*Drosophila melanogaster*)	*Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae*, *Drosophila pseudoobscura*, *Drosophila persimilis*, *Drosophila willistoni*, *Drosophila virilis*, *Drosophila mojavensis*, *Drosophila grimshawi*, *Anopheles gambiae*, *Anopheles mellifera* and *Tribolium castaneum*
Yeast (*Saccharomyces cerevisiae*)	*Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus*, *Saccharomyces castelli*, and *Saccharomyces kluyveri*

web server can display the distributions of indels in different genomic regions [including coding sequence (CDS), untranslated region (UTR), intron and intergenic region]. The genomic sequences and annotations of 9 target species and more than 35 subject/out-group species (see Table 1) are available for the current version of INDELSCAN for analysis. The server also provides the target-subject pair-wise indels for comparison.

MATERIALS AND METHODS

Process flow of INDELSCAN

The system flow of TSS/NSS indel identification and categorization are stated below. First, the user can upload multiple sequence alignments of at least three species in the UCSC (University of California, Santa Cruz) multiple alignment format (described at http://genome.ucsc.edu/goldenPath/help/maf.html) and specify the target, subject and out-group species. The user can also select the compared species from the INDELSCAN-provided species list, which is linked to pre-stored genomic sequences downloaded from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/downloads.html). To reduce potential errors, only indels that occur within continuously alignable sequences are considered. Second, overlapping alignments (i.e. one target sequence segment is aligned to two or more genomic sequences in the compared species) are filtered out in the system to eliminate potential spurious indels. Third, three indel types (also see Figure 1), TSS (Events 1 and 2), subject species-specific (‘SSS’; Events 3 and 4) and NSS (Events 5 and 6) indels, are identified. The TSS indels are classified into TSS insertions and TSS deletions based on comparison with the out-group genomic sequence(s) (described below). Fourth, using the user-input or INDELSCAN-provided annotations of the target genome, the genomic locations (i.e. 5′UTR, intron, CDS, 3′UTR and intergenic region) of the identified indels are determined. Finally, the system adjusts the UCSC pair-wise sequence alignments (described below) and provides pair-wise indels between the target and subject genomes for comparison. Analytic figures/tables that compare the numbers and rates of TSS, SSS, NSS and pair-wise indels are also provided.

Discrimination between TSS insertions and TSS deletions

As stated above (also see Figure 1), the inclusion of out-group genomic sequences enables the system to distinguish between insertions and deletions in TSS indels. Here, a TSS insertion is defined as a DNA segment (or a single base) of the target species genome that is not only absent in the orthologous genomic sequence of the subject species but also absent or partially absent in the out-group species (e.g. Event 1). A TSS deletion is defined in a similar way (e.g. Event 2). If the subject species genome compared is still in the draft stage, then many indels (especially one- or two-bp indels) inferred from target-subject pair-wise sequence alignments may be false positives that result from sequencing or assembling errors. The TSS indels identified by INDELSCAN are therefore more reliable than indels inferred from pair-wise comparisons because the former are supported by the genomic sequences of the out-group species.

Update of UCSC multiple sequence alignments of vertebrate genomes

The chimpanzee genome used in the current UCSC multiple alignments of vertebrate genomes is Build 1 Version 1 (or UCSC version panTro1). To include the up-to-date chimpanzee genome (Build 2 Version 1 or UCSC version panTro2), three processes were performed. First, in the UCSC human-genome-based (hg18) multiple alignments, we replaced panTro1 with panTro2 transplanted from the UCSC human–chimpanzee pair-wise alignments (hg18 versus panTro2). Second, we used the MUSCLE package (11,12) to realign the updated sequences. Finally, the new alignments were transformed to the UCSC multiple alignment format.

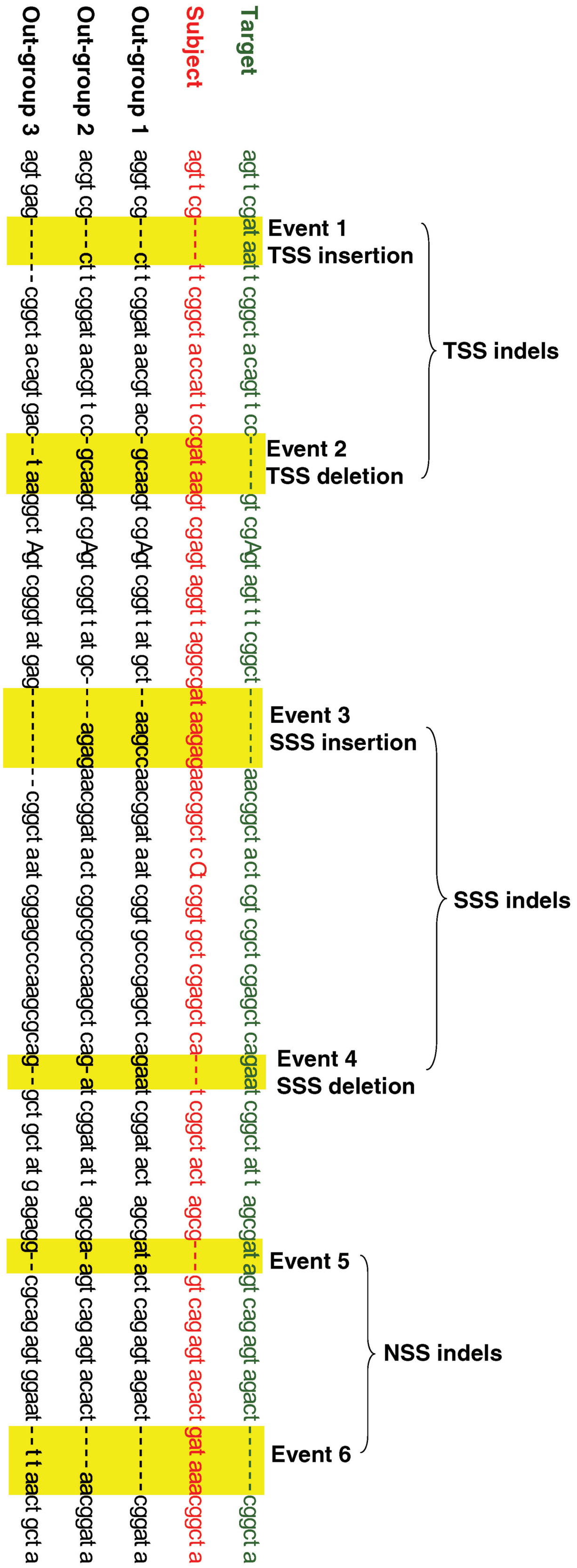

Target
Subject
Out-group 1
Out-group 2
Out-group 3
agt t cgat aat t cggct acagt t cc- - - - gt cgAgt agt t t cggct - - - - - aacggct act cgt cgct cgagct cagat cggct at t agcgat agt cag agt agact - - - - - cggct a
agt t cg- - - t t cggct accat t ccgat aagt cgagt aggt t aggcgat aagagaacggct cCt cggt gct cgagct ca- - t cggct act agcg- - gt cagagt acact gat aaacggct a
aggt cg- - - ct t cggat aacgt acc-gcaagt cgAgt cggt t at gct - - aagccaacggat aat cggt gcccgagct cagaat cggat act agcgat act cagagt agact - - - - - - cggat a
acgt cg- - - ct t cggat aacgt t cc-gcaagt cgAgt cggt t at gc- - - agagaacggat act cggcgcccaagct cag-at cggat at t agcga-agt cagagt acact - - - aacggat a
agt gag- - - - - cggct acagt gac- - t aaggct Agt cggt at gag- - - - - - - - cggct aat cggagcccaagcgcag- - gct gct at g agagg- - cgcagagt ggaat - - t t aact gct a
Event 1
TSS insertion
Event 2
TSS deletion
TSS indels
Event 3
SSS insertion
Event 4
SSS deletion
SSS indels
Event 5
Event 6
NSS indels

Figure 1. Three types of INDELSCAN-identified indels: TSS indels (Events 1 and 2), SSS indels (Events 3 and 4) and NSS indels (Events 5 and 6), with odd number events representing insertions and even number events, deletions.

Justification of pair-wise sequence alignments

In the UCSC alignments, there exist a considerable number of potentially artificial alignment gaps,

e.g.
aagcatgcat—gaatcggata
aagcatg—agttgaatcggata

Such alignments might result in false indel inferences. To eliminate these gaps, we realigned and closed neighboring UCSC alignment gap pairs that satisfied both of these conditions: (i) one gap occurs in the subject genome and the other in the target genome and (ii) the distance between these two gaps is not larger than three bases. And the adjustment process continues until no gap pairs satisfy the conditions. The alignment shown above is adjusted as follows:

aagcatgcat-gaatcggata
aagcatgagttgaatcggata

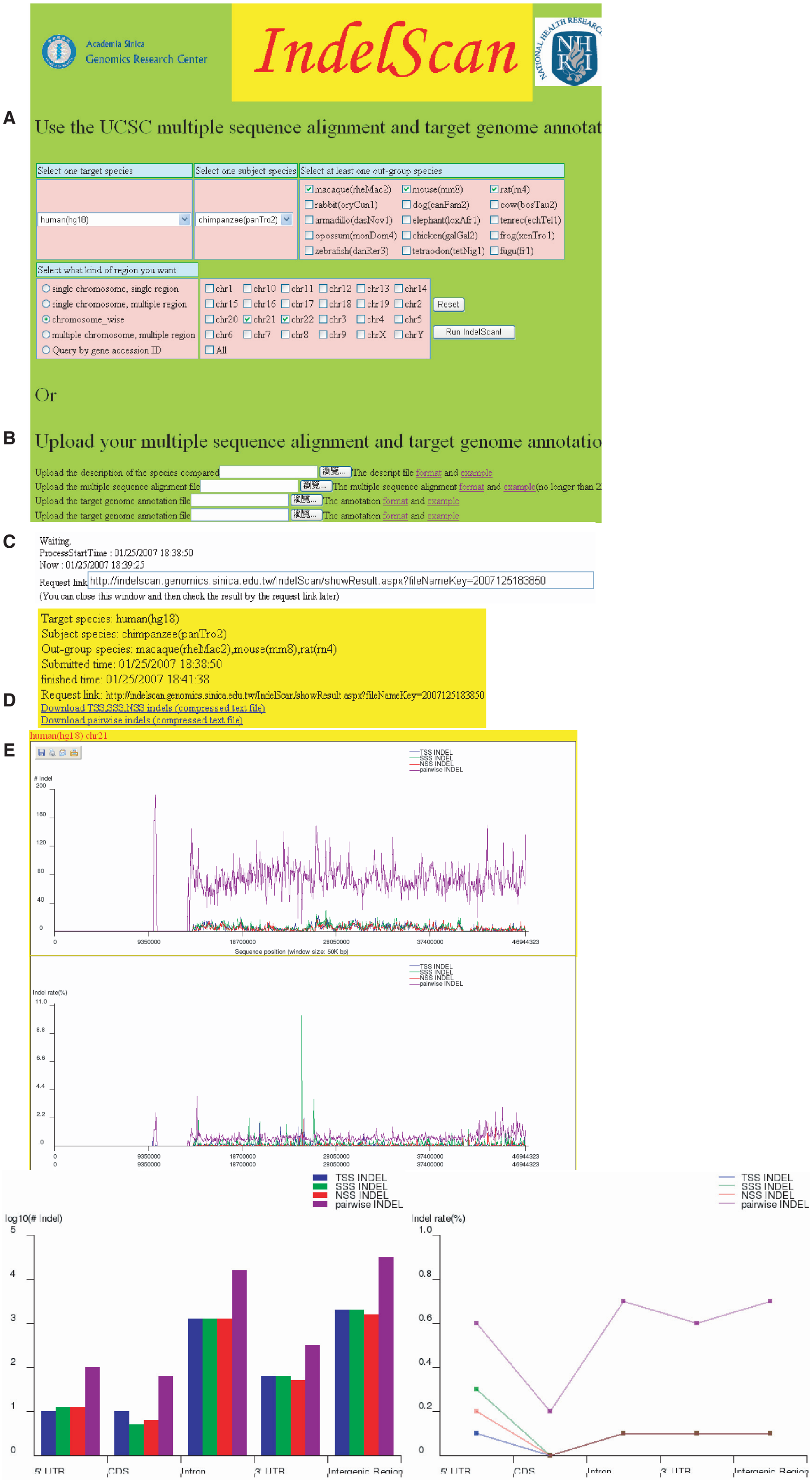
Implementation and run time

Our server is implemented in ASP.NET on the server end and java script on the client end. There are four major steps in this program: scanning the multiple sequence alignments and filtering out overlapping alignments; inferring TSS, SSS and NSS indels from the multiple sequence alignments; determining the genomic locations of the identified indels and identifying target-subject pair-wise indels. As an example of run time, analysis of indels in the human chromosomes 21 and 22 using human as target, chimpanzee as subject and macaque, mouse and rat as out-group species takes approximately 3 min (see Figure 2).

WEB SEVER DESCRIPTION

Input

INDELSCAN supports two schemes for inputting multiple alignments and annotation information in the target genome: users can use the INDELSCAN-provided alignments/annotation (Figure 2A) or upload their own data (Figure 2B). For the former scheme, users can select one target, one subject and at least one out-group species in the system. The web server allows users to choose one out of nine target species, including human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*), opossum (*Monodelphis domestica*), chicken (*Gallus gallus*), zebrafish (*Danio rerio*), fly (*Drosophila melanogaster*) and yeast (*Saccharomyces cerevisiae*). Table 1 lists the available subject and out-group species for each target species. As shown in Figure 2A, users can also choose to perform indel analysis for single or multiple chromosomes or only in user-specified region(s)/gene(s). For the latter scheme, users must upload three files to the system: description of compared species, multiple sequence alignments and target genome annotation (Figure 2B). Note that INDELSCAN by default takes the first sequence in the uploaded multiple alignments as target species sequence, and the second as subject-species sequence, whereas the others are regarded as out-group sequences. Also note



Use the UCSC multiple sequence alignment and target genome annotat

Select one target species: human(hg18)

Select one subject species: chimpanzee(panTro2)

Select at least one out-group species

☑ macaque(rheMac2) ☑ mouse(mm8) ☑ rat(rn4)
☐ rabbit(oryCun1) ☐ dog(canFam2) ☐ cow(bosTau2)
☐ armadillo(dasNov1) ☐ elephant(loxAfr1) ☐ tenrec(echTel1)
☐ opossum(monDom4) ☐ chicken(galGal2) ☐ frog(xenTro1)
☐ zebrafish(danRer3) ☐ tetraodon(tetNig1) ☐ fugu(fr1)

Select what kind of region you want:

○ single chromosome, single region
○ single chromosome, multiple region
● chromosome_wise
○ multiple chromosome, multiple region
○ Query by gene accession ID

☐ chr1 ☐ chr10 ☐ chr11 ☐ chr12 ☐ chr13 ☐ chr14
☐ chr15 ☐ chr16 ☐ chr17 ☐ chr18 ☐ chr19 ☐ chr2
☐ chr20 ☑ chr21 ☑ chr22 ☐ chr3 ☐ chr4 ☐ chr5
☐ chr6 ☐ chr7 ☐ chr8 ☐ chr9 ☐ chrX ☐ chrY
☐ All

Or

Upload your multiple sequence alignment and target genome annotatio

Upload the description of the species compared The descript file format and example
Upload the multiple sequence alignment file The multiple sequence alignment format and example(no longer than 2
Upload the target genome annotation file The annotation format and example
Upload the target genome annotation file The annotation format and example

Waiting.
ProcessStartTime : 01/25/2007 18:38:50
Now : 01/25/2007 18:39:25
Request link http://indelscan.genomics.sinica.edu.tw/IndelScan/showResult.aspx?fileNameKey=2007125183850
(You can close this window and then check the result by the request link later)

Target species: human(hg18)
Subject species: chimpanzee(panTro2)
Out-group species: macaque(rheMac2),mouse(mm8),rat(rn4)
Submitted time: 01/25/2007 18:38:50
finished time: 01/25/2007 18:41:38
Request link: http://indelscan.genomics.sinica.edu.tw/IndelScan/showResult.aspx?fileNameKey=2007125183850
Download TSS,SSS,NSS indels (compressed text file)
Download pairwise indels (compressed text file)

Figure 2. The INDELSCAN web sever. **(A)** Users can choose from pre-stored genomic sequences for comparison or **(B)** upload their own multiple sequence alignments and annotations. **(C)** The server will exhibit the time elapsing and request link after the job is submitted. **(D)** When the job is completed, the server gives the links from which the results can be downloaded. **(E)** Also provided are analytical figures that show the numbers of indels, indel rates and genomic distributions of indels (results in the human chromosome 22 are not shown in the figure).

that only the species specified in the description file will be processed in the server. For each submitted job, INDELSCAN will automatically assign a request link for the user to retrieve the results (Figure 2C).

Output

When the submitted job is completed, users can use the request link to visualize or download their results. Two text outputs (indels inferred from multiple sequence alignments and pair-wise indels) are also downloadable from the system (Figure 2D), including the coordinates of the identified indels in the target and subject genomes, indel lengths, genomic locations of indels and the IDs of indel-affected genes. The result page also shows analytic figures for comparisons of the numbers and rates of TSS, SSS, NSS and pair-wise indels for each user-specified region (Figure 2E). The results of each query will be retained in the system for 48 h.

CONCLUSION

The INDELSCAN web interface identifies TSS, SSS and NSS indels using multiple sequence alignments. So far, such comparisons have remained scarce due to lack of suitable analysis tools and high-quality genomic sequences. With the rapidly increasing number of available genomes, multi-genome comparisons will soon become a norm. INDELSCAN will therefore, be helpful for analysis of species-specific and non-species-specific indels. Moreover, the server also detects the genomic locations of the identified indels, giving very useful information for biologists to functionally study these indels. Note that the web interface can identify indels on a genome-wide scale or in individual genes or shorter sequences of interest. Individual genes and sporadic sequences are more readily accessible than complete genome sequences. It is relatively easy to compare shorter homologous sequences of a large number of species for inference of species specificity, which would otherwise be limited to only a few compared species in genome-scale comparisons. Such large-number-species comparisons can provide high resolution for inference of species specificity and paths of indel evolution through the phylogenetic tree of the compared species. Currently, the INDELSCAN-provided multiple sequence alignments are downloaded from the UCSC genome browser, which deals with gaps by assigning gaps to branches in the phylogenetic tree using out-group information (13). For multiple sequence alignment tools, to measure the cost of a multiple alignment and to choose gap costs consistent with the measure chosen remain challenging (14–16). Many tools have been proposed and have performed well. For example, CLUSTALW (17) dynamically adjusts the gap penalties in a position- and residue-specific manner. T-Coffee (18) improves the accuracy but sacrifices the efficiency by first building a library of both local and global alignments for each pair of sequences and then using a library-based scoring scheme for progressive alignment. MUSCLE (11,12) and MAFFT (19) enhance both the accuracy and efficiency by initially building a progressive alignment and then horizontally refining the phylogenetic tree to improve the objective score. In our server, indel analyses will not be limited to the sequence alignments available from UCSC, but can be performed together with any multiple sequence alignment tools. Moreover, INDELSCAN can be applied to the detection of lineage-specific indels, such as primate-, rodent-, mammal-, avian- or fish-specific indels. Results thus obtained may bring functional and evolutionary insights and help focus experimental studies. Finally, analyses of indel events can be combined with other species-specific events, such as nucleotide substitutions, frame-shift mutations (20), duplications (21) and pseudogenizations (22), to further our understanding of the mechanisms of speciation and functional divergence.

ACKNOWLEDGEMENTS

We thank Yao-Ting Huang for comments on the web server. This work is supported by the Genomics Research Center (GRC), Academia Sinica, Taiwan and the National Health Research Institutes (NHRI), Taiwan, under contract NHRI-EX95-9408PC (to TJC) and NHRI intramural grant (to FCC). Funding to pay the Open Access publication charge was provided by GRC, Academia Sinica, Taiwan (to TJC).

Conflict of interest statement. None declared.

REFERENCES

1. Britten,R.J. (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science*, **231**, 1393–1398.
2. Frazer,K.A., Chen,X., Hinds,D.A., Pant,P.V., Patil,N. and Cox,D.R. (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.*, **13**, 341–346.
3. Watanabe,H., Fujiyama,A., Hattori,M., Taylor,T.D., Toyoda,A., Kuroki,Y., Noguchi,H., BenKahla,A., Lehrach,H. *et al.* (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, **429**, 382–388.
4. Anzai,T., Shiina,T., Kimura,N., Yanagiya,K., Kohara,S., Shigenari,A., Yamagata,T., Kulski,J.K., Naruse,T.K. *et al.* (2003) Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proc. Natl Acad. Sci. USA*, **100**, 7708–7713.
5. Chen,F.C., Chen,C.J., Li,W.H. and Chuang,T.J. (2006) Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.*, **17**, 16–22.
6. The Chimpanzee Genome Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
7. Newman,T.L., Tuzun,E., Morrison,V.A., Hayden,K.E., Ventura,M., McGrath,S.D., Rocchi,M. and Eichler,E.E. (2005) A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.*, **15**, 1344–1356.
8. Stedman,H.H., Kozyak,B.W., Nelson,A., Thesier,D.M., Su,L.T., Low,D.W., Bridges,C.R., Shrager,J.B., Minugh-Purvis,N. *et al.* (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*, **428**, 415–418.
9. Hayakawa,T., Satta,Y., Gagneux,P., Varki,A. and Takahata,N. (2001) Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. *Proc. Natl Acad. Sci. USA*, **98**, 11399–11404.
10. Hamann,J., Kwakkenbos,M.J., de Jong,E.C., Heus,H., Olsen,A.S. and van Lier,R.A. (2003) Inactivation of the EGF-TM7 receptor

EMR4 after the Pan-Homo divergence. *Eur. J. Immunol.*, **33**, 1365–1371.

11. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
12. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
13. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
14. Altschul,S.F. (1989) Gap costs for multiple sequence alignment. *J. Theor. Biol.*, **138**, 297–309.
15. Lipman,D.J., Altschul,S.F. and Kececioglu,J.D. (1989) A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA*, **86**, 4412–4415.
16. Chan,S.C., Wong,A.K. and Chiu,D.K. (1992) A survey of multiple sequence comparison methods. *Bull. Math. Biol.*, **54**, 563–598.
17. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
18. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
19. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
20. Hahn,Y. and Lee,B. (2005) Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics*, **21**(Suppl. 1), i186–i194.
21. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
22. Wang,X., Grus,W.E. and Zhang,J. (2006) Gene losses during human origins. *PLoS Biol.*, **4**, e52.